基于视觉的人体行为识别算法研究综述 *

陈煜平，邱卫根

(广东工业大学 计算机学院，广州 510006)

摘　要：人体行为识别应用广泛，是人工智能领域研究的热点问题，针对人体行为识别算法进行归纳总结，具有很重要的参考价值。以行为识别为核心，同时包含数据集、动作分割等内容。引言部分主要讲述人体行为识别的基础流程，数据集部分归纳了人体行为识别常用的数据集，动作分割方法总结了时域分割的发展现状和常用的方法，传统方法讲解了人体行为识别比较经典的方法，深度学习方法归纳了人体行为识别最新最热的深度学习方法。引入了动作分割，再结合行为识别，能够实现连续的人体行为识别，使得行为识别适用于实际场景，而不再是对经过人工剪辑好的单个视频进行识别，这在实际应用中意义重大。

关键词：人体行为识别；数据集；动作分割；深度学习；双流网络

中图分类号：TP391.41　　doi: 10.3969/j.issn.1001-3695.2018.04.0259

Survey of human action recognition algorithms based on vision

Chen Yuping, Qiu Weigen

(*School of Computers Guangdong University of Technology*, *Guangzhou* 510006, *China*)

Abstract: Human action recognition is a hot issue in the field of artificial intelligence. So it has important reference value to summarize the human action recognition algorithms. This paper focused on action recognition and included data sets and motion segmentation. The introductory part mainly described the basic flow of human action recognition. And the data sets part summarized the commonly used data sets of human action recognition. Then the motion segmentation method summarized the development status and common methods of time domain segmentation. Next the traditional methods explained the classic algorithms of human action recognition. At last, the deep learning methods summarized the the-state-of-art deep learning methods of human action recognition. The introduction of action recognition combines with action segmentation makes the action recognition applicable to the actual scene, which can achieves continuous recognition of human action. Meanwhile it is no longer recognize a single video that has been manually edited. This has very important reference value in practical applications.

Key words: human action recognition; data set; motion segmentation; deep learning; two-stream network

0　引言

人体行为识别主要根据采集到的视频来分析人体行为，这在视频监控、医疗康复、健身评估、人机交互等领域应用广泛，是计算机视觉研究的热点问题。

从实现方式来分类，可以把人体行为识别分为基于传感器和视觉两种，当然也包含这两者的结合。基于传感器的行为识别由于要佩戴相应的传感器，使用不够灵活，操作复杂，扩展性不强，用户体验得不到有效保证等原因，因此只能在一些特定领域中使用；基于视觉的行为识别又可以分为基于单帧图像和视频的识别，基于单帧图像的行为识别由于不能有效获取行为的连贯时间信息，通常会产生误判，而基于视频的行为识别能够很好的获取视频中的空间和时间信息，准确率得到大大的提升。基于视频的行为分析由于扩展性强，灵活度高，得到了广泛的研究和应用。

人体行为识别处理流程一般可以分为特征提取、特征处理、学习算法输出结果三步。首先从原始视频中提取特征，经过一定的处理，形成一个特征描述符，最后通过学习算法实现分类。对于有些学习算法输入维度固定，而特征描述符不固定的情况，还要通过一定的方法对特征描述符进行聚合，使得输入维度固定。本文加入了动作分割，即可实现连续的人体行为识别，流程图如图 1 所示。

本文主要从数据集、动作分割、 传统方法和深度学习方法四个方面进行介绍。首先介绍目前常用的人体行为识别数据库，

收稿日期：2018-04-30；**修回日期：**2018-06-10　**基金项目：**国家自然科学基金资助项目（61572142）；广东省科技计划资助项目（14ZK0180）

作者简介：陈煜平（1992-），男，江西安福人，硕士，主要研究方向为计算机视觉、图像处理、深度学习（1322633925@qq.com）；邱卫根（1968-），男，江西临川人，教授，博士，主要研究方向为人工智能、粗糙集理论及应用、计算机图形图像学.

包含二维和三维数据库。然后介绍动作分割，这主要针对连续的行为识别，相对于传统的只支持单个动作的识别有更大的优势。传统方法主要介绍行为识别的常用方法，而深度学习方法主要介绍最近几年基于深度学习的行为识别技术。

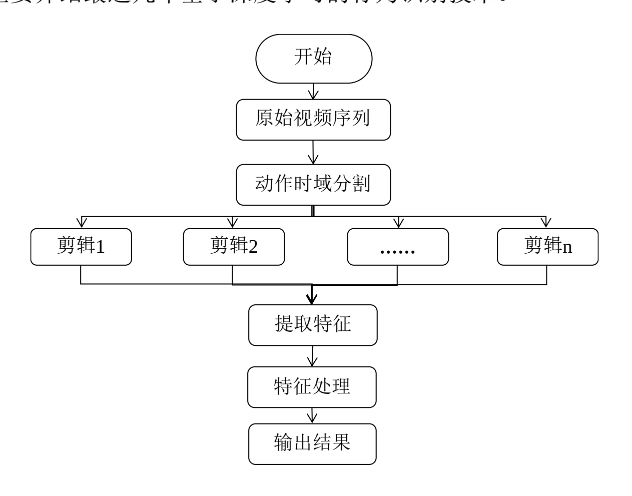


图 1　行为识别基本流程

1　数据集

为了方便实验，出现了很多人体行为识别数据集[1]，可以把它划分成二维和三维数据集，二维数据集一般用普通的摄像头进行采集，而三维数据集一般用如 Kinect 等可以获取深度信息的特殊摄像头进行采集，三维数据集由于包含图像的深度信息，因此信息更加丰富。

1.1　二维数据库

二维数据库起步较早，从发展趋势来看，数据集逐步趋于复杂，行为种类更加丰富，场景更加多样，每个行为的样本数变得愈加庞大，对算法提出了更加严苛的挑战。表 1 列出了常用的二维数据集。

表 1　常用二维数据集概览

数据集	时间	行为类别	视频数	数据来源
KTH[2]	2004	6	599	实验采集
Weizmann[3~4]	2005	10	93	实验采集
IXMAS[5]	2006	13	180	实验采集
Hollywood[6]	2008	8	663	电影
Hollywood 2[7]	2009	12	3669	电影
UCF sports[8]	2008	10	150	广播电视
UCF YouTube[9]	2008	11	1600	YouTube
UCF50[10]	2013	50	6676	YouTube
UCF101[11]	2012	101	13320	YouTube
HMDB51[12]	2011	51	6849	电影和其他公共资源
Sports-1M[13]	2014	487	1133158	YouTube

KTH[2]是最早的人体行为数据集，它包含的数据比较少，只有 6 类行为共 2391 个视频，一共 25 个人的数据，虽然该数据集比较简单，但是对人体行为识别起到了里程碑式的作用。Weizmann[3,4]数据集包含 10 类行为共 93 个视频，一共 9 个人的数据，也是比较小的一个数据集。IXMAS[5]包含 13 种行为共 180 个视频，虽然该数据集比较小，但是包含了 5 个视角，对多视角的研究提供了比较可靠的数据。这三个数据集都是早期比较通用的数据集，数据量少，场景简单，没有复杂背景，目前很少使用。

Hollywood[6]的数据来自于好莱坞电影中的视频剪辑，从 32 部电影中剪辑出了 8 类行为共 663 个视频。Hollywood 2[7]是在 Hollywood 的基础上进行扩展，从 69 部电影中剪辑出了 12 类行为共 3669 个视频。由于这两个数据集接近真实场景，包含复杂的背景、光照变化、自遮挡等，具有一定的挑战性。

UCF 包含了一系列数据集，种类繁多，受到了广泛关注，富有挑战性。UCF sports[8]的数据来源于 BBC 和 ESPN 等电视频道，包含 10 类行为共 150 个视频，它的视频分辨率比较高，接近自然场景。UCF YouTube[9]现称 UCF11，数据来源于 YouTube，包含 11 类行为共 1600 个视频，它对视频进行了分组，每组具有一些共同特性。UCF50[10]的数据来源于 YouTube，行为类别从 UCF YouTube 的 11 种扩展到了 50 种，共包含 6676 个视频，具有一定的挑战性。UCF101[11]是对 UCF50 的扩展，包含 110 种动作类别共 13320 个视频，由于该数据集包含很多低质量和不同光照的视频，因此极具挑战性。

HMDB51[12]数据主要来源于电影和一些公共资源，包含 51 种行为类别共 6849 个视频，由于该数据集来源多样，视频中包含的场景复杂、光照条件变化等因素，是目前最具挑战性的数据集之一。Sports-1M[13]的数据集来自于 YouTube，是 2014 年 Google 公布的一个大型数集，包含 289 种行为类别共 1133158 个视频，每个动作类别包含 1000 到 3000 个视频，超过一百万个视频的庞大数据集，是前面的数据集无法逾越的优势，该数据集包含场景多，种类多样，极具挑战性。

1.2　三维数据集

由于人体行为存在自遮挡等问题，二维的数据不能很好地解决这些问题，而三维的数据能够提供更多的信息，对自遮挡的信息得到了补充，使得人体行为识别变得相对容易，但会使得数据集变得复杂，处理起来也变得相对困难。然而由于计算机硬件的发展，三维数据的采集和处理变得容易和方便，微软 Kinect 的应用就是一个很好的例子。表 2 列出了常用的三维数据集。

表 2　三维数据集概览

数据集	时间	关节点数	动作类别	样本数	采集设备
CMU Motion Caption(Mocap)[14]	2016	41	6	2605	8 个红外摄像头
MSR Action 3D[15]	2010	20	20	567	Kinect v1
MSR Daily Activity 3D[16]	2012	20	10	320	Kinect v1
UCF Kinect[17]	2013	15	16	1280	Kinect v1
N-UCLA Multiview Action3D[18]	2014	20	10	1493	三台 Kinect v1
UTD-MHAD[19]	2015	20	17	861	Kinect v1 和 IMU
NTU RDB+D[20]	2016	25	60	56880	Kinect v2

CMU Motion Caption（Mocap）[14]是卡内基梅隆大学发布

的一个三维数据集，通过 8 个红外摄像头进行采集，并提供了 41 个人体关节点信息，它包含 6 大类行为和 23 个子类行为共 2605 个数据，每个大类数据又包含一个或多个行为类别，提供的数据能够构建完整的人体三维行为模型。

由于微软 Kinect 的出现，很多数据库在此基础上进行了采集工作。MSR Action 3D[15]使用 Kinect 进行采集，提供 20 个节点的人体骨架数据和深度图，包含 20 种行为共 567 个数据，该数据集的视频比较纯粹，没有背景，但由于噪声等原因，依然有一定的难度，目前使用比较广泛。MSR Daily Activity 3D[16]也是通过 Kinect 进行采集，主要是生活中的日常行为，包含 10 种行为共 320 个数据，该数据集背景是真实环境，所以更具挑战性。UCF Kinect[17]同样采用 Kinect 采集数据，但是它没有用微软的 SDK 来评估骨架序列，而是用 OpenNI 来评估骨架序列，每个序列有 15 个骨架节点，包含 16 种行为共 1280 个数据。N-UCLA Multiview Action3D[18]数据集采用了三台 Kinect 进行采集，因此包含了三个视角，包含 10 种行为共 1493 个数据，该数据集的每个行为都是从不同视角采集的，具有一定的挑战性。UTD-MHAD[19]通过 Kinect 和 IMU 进行采集，提供 20 个骨架节点，包含 17 种行为共 861 个数据。NTU RGB+D[20]通过第二代 Kinect 进行采集，提供 25 个骨架点，包含 60 种行为共 56880 个数据。

2 动作分割

此处的动作分割指的是把连续的动作从视频中分割出来，即时域分割，也就是说如果一个视频中包含走、跑、跳等动作，动作分割算法能准确判断每个动作的的边界，并把该动作从原视频中分割出来。由于目前行为识别都是在已经分割好的数据集上进行的，而在现实中采集的数据都是未进行动作分割的视频，因此动作分割对实现连续的人体行为识别至关重要。

2.1 基于 PCA 的方法

Barbič 等人[21]提出了三种方法，PCA 方法、PPCA 方法和 GMM 方法，PCA 方法基于这样的想法：包含单个行为的运动序列的固有维数应该小于包含多个行为的运动序列的固有维度。通过计算离散的误差 *di*，当 d_i 急剧上升超过固定值时，就认为这是过渡点。PPCA 是在 PCA 的基础上改进而来，基于假设“动作序列符合高斯分布，两个不同的动作将会有很大的区别”，采用滑动窗口（同时前向滑动和后向滑动）的机制来找过度点，通过计算滑动窗口的马氏距离，当到达极大值点时，认为这个是过渡点。GMM 基于假设“动作序列中的每一个动作都符合不同的高斯分布”，通过 PCA 投影在超平面上，用 EM（expectation maximization）来评估高斯模型的参数，以此达到分割的目的。

基于 PCA 的方法都基于一定的假设，具有一定的局限性，其中 GMM 方法还要求事先知道每个视频中包含的动作类别数，而大多数情况下是未知的。不过该类方法对硬件的要求不高，实现也相对比较简单，对一些符合假设的应用可以轻易实现。

2.2 基于聚类的方法

Zhou 等人[22]提出了基于聚类方法的动作分割 ACA（Aligned Cluster Analysis），ACA 使用两种方式扩展了标准的 kmeans 聚类：a）聚类包含可变数量的特征；b）动态时间规整（DTW）内核用于实现时间不变性。使用了 DTAK（dynamic time alignment kernel）来对两个时间序列进行度量，因为 DTW 不是一个正确定义的度量，它不满足三角形约束。Zhou 等人[23]提出了 HACA 方法，是在 ACA 的基础上进行了改进，HACA 提供了用于聚类和可视化时间序列数据的自然嵌入，并提供了几个时间尺度的分层分解。时间聚类问题是能量最小化，而最小化 HACA 是一个 NP 问题，通过动态规划提出了一种有效的坐标下降最小化方法。

Xia 等人[24]提出了基于 SSC（sparse subspace clustering，稀疏子空间聚类）的方法，通过 SSC 进行子空间聚类，再使用三角形约束解决在不同的时间段内相似帧不会被分到同一个簇，保证了时间上的连续性，并通过相关熵抑制子空间聚类的非高斯噪声。最后所有的系数矩阵的绝对平均值将用于最终的分割相似矩阵，目的是重建不同关节点之间的关系，而不会因为把整个序列视为一个整体而忽略关节点之间的关系。

基于聚类的方法一般整体效果要优于基于 PCA 的方法，时间复杂度也相对较高，但对硬件要求不高，综合成本和需求考虑，是作为应用的理想选择。

2.3 基于深度学习的方法

Lea 等人[25]提出了 TS-CNN 方法，引入了低级编码视觉信息的时空 CNN（ST-CNN）和捕获高级时间信息的半马尔可夫模型。ST-CNN 的空间分量是 VGG 的一个变体，用于编码对象状态、位置和对象间关系的细粒度任务。分段组件使用半马尔可夫和条件随机场（CRF）共同分割并分类动作。

Lea 等人[26]提出了 TCN 方法，如图 2 所示，TCN 又可以分为 ED-TCN 和 Dilated TCN，其中 ED-TCN 引入了编码和解码网络，而 Dilated TCN 是从 WaveNet 改进过来的，但他们又具有共同的特点：a）计算层次执行，这意味着每个时间步同步更新，而不是逐帧更新；b）卷积是跨时间计算；c）在每个帧处的预测是固定长度的时间段的函数，其被称为接收场。其中 ED-TCN 的效果优于 Dilated TCN。

该类方法是目前研究的热点，一般都使用 CNN（或自动编码器）加其他机器学习方法的组合，效果也要优于其他方法，但是对硬件配置的要求较高，而且依赖大量的数据，实现起来比较困难，不过模型小型化也是一个不错的选择。

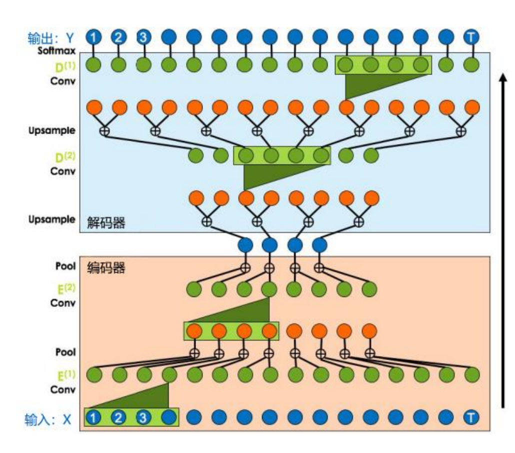
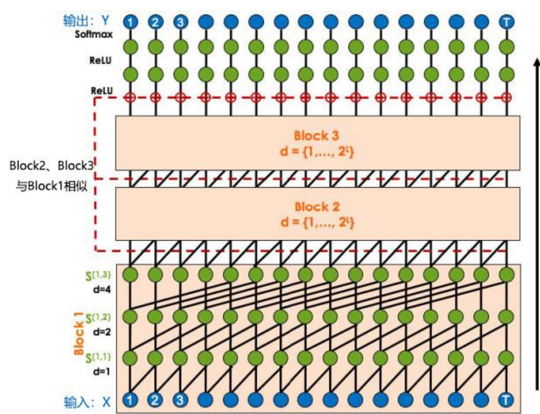



图 2　Lea C 等人[25]提出的 TCN 方法,左图为 ED-TCN 的结构,右图为 Dilated TCN 的结构

2.4　其他方法

Devanne M 等人[27]在黎曼形状空间中研究运动轨迹形状，并通过动态朴素贝叶斯分类器实现动作分割。Vögele A 等人[28]使用邻域图的方法来分割动作序列，Borzeshi E Z 等人[29]使用了扩展的 HMM（HMM-MIO）模型来进行动作分割和识别，其中分割和分类是放在一起进行的。Liu S 等人[30]提出了 TS-WMCS 方法，利用时间序列扭曲的曲率实现分割。

3　传统方法

传统方法主要是人工提取特征，还要建立起表示人体行为的模型，再在建立好的模型上识别人体行为，可以从表示方式上分为整体表示方法和局部表示方法。

3.1　整体表示方法

整体表示方法把人体行为表示为一个整体，用于整体分析人体行为。Bobick[31]提出了经典的 MEI（运动能量图像）和 MHI（运动历史图像），其基本思想是通过一个图像对运动相关信息进行编码。其中 MHI 模板显示运动图像如何移动，MHI 中的每个像素是该点处的运动的时间历史的函数（即更高的强度对应于更近的运动），因此 MEI 和 MHI 模板包含有关视频上下文的有用信息。Blank 等人[32]提出了 MEI 模板的体积扩展，主要思想是通过在时空中的剪影引起的三维形状来表现动作，如图 3 所示。Weinland D 等人的研究[33]建议通过时空体积代表 MHI 模板，并表明三维体积的扩展增加了视点变化的鲁棒性。



图 3　Blank M 等人[32]提出的三维形状来表现行为为动作

Yilmaz A 等人[34]根据时空体积（STV）的不同特性来确定行为。STV 是通过沿时间轴叠加物体轮廓而建立的。STV 的方向，速度和形状的变化表征了潜在的动作。动作描述是从 STV 的表面提取的一组属性（如高斯曲率），并且对于观察点变化显示是鲁棒的。

整体方法也存在一定的不足。Dollar 等人[35]的研究表明，整体的方法太僵化，不能有效捕捉行为的视点、遮挡等变化，Matikainen P 等人[36]认为，基于轮廓的表示不能捕捉轮廓内的细节。因此，目前局部表示方法和深度特征更受青睐。

3.2　局部表示方法

局部表示方法把视频中的一个局部区域用来描述人体的运动，Laptev 等人[37]提出的时空兴趣点（STIPs）为局部表示方法的人体行为识别奠定了基础，行为识别的局部表示遵循三个流程：兴趣点检测 、局部描述符提取、局部描述符聚合。

兴趣点检测：提取时空兴趣点，需要构建 STIP 探测器，而构建探测器也有多种方法。Laptev 等人[38]将 Harris 角点探测器[39]扩展到 3D-Harris 探测器，在 3D Harris 中，除了丰富的空间结构之外，还需要时间上的重要性来激发探测器，3D Harris 探测器识别具有大空间变化和非恒定运动的点。为了解决相机摇晃激发的一些列不相关的兴趣点，Liu J 等人[39]建议使用检测到的兴趣点的统计属性修剪不相关的特征。当然，还有很多其他的方法和在上述方法的基础上改进的方法，但最经典的还是 Harris 角点探测器。

局部描述符提取：提取时空兴趣点之后，还得对兴趣点进行处理，形成一个局部的描述符来描述人体行为。Kläser 等人[40]建议使用梯度方向直方图作为运动描述符，描述符本身受到面向方向梯度直方图（HoG）[41]的启发，其本身跨越时空域，因此被命名为 HoG3D 描述符。Laptev 等人[42]将局部区域上的光流直方图（HoF）作为时空描述符，HoF 描述符更鲁棒的扩展是 Dalal 等人[43]引入的运动边界直方图（motion boundary histogram，MBH）。局部二进制模式（LBP）是基于强度的二维描述符，成功地用于包括人脸识别和纹理分析在内的多种视觉问题[44]，LBP 描述符通过量化关于其强度的像素的邻域来计算。Zhao 等人[45]介绍了二维 LBP 描述符到时空域的各种扩展 volume LBP（VLBP），局部 volume 由二进制模式的直方图编码，LBP 还有一些变体[46~47]。Sanin 等人[48]提出了通过二阶统计描述图像区域。

时空兴趣点可能并不位于长方体的时间延伸内完全相同的空间位置处，因此从长方体提取的特征可能不一定描述兴趣点本身。轨迹是随着时间的推移正确跟踪的特征，如图 4 所示，轨迹提取局部特征主要由 Messing 等人[49]和 Matikainen 等人[36]

提出，他们都使用轨迹速度作为局部特征。Jiang 等人[50]和 Wang 等人[51]使用相机运动校正轨迹来改进前面的轨迹。

局部描述符的选择，最受青睐的当属轨迹以及轨迹的一些改进方法，而对于兴趣点是选取稀疏还是密集，Wang H 等人[52]进行了详细的比较，一般而言，密集的效果会优于稀疏，但时间和空间复杂度会相对较高。

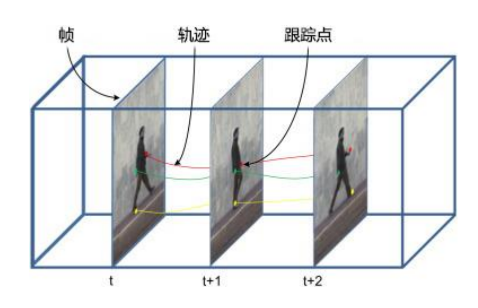


图 4　轨迹描述符的形成

局部描述符聚合：从视频中提取局部特征后，为了能够对特征进行处理，通常会使用诸如 SVM 等学习算法来训练得到最终结果，但通常这类算法大多数只接受固定大小的向量输入，所以需要一种机制来使得提取的局部特征聚合成固定大小的描述符，主要有三种机制。第一种主要通过 BoV，简言之就是通过给定一个“视觉词袋”或“码本”，局部描述符在码本上的分布用作描述符，相关的工作主要有 Dollar P 等人[35, 40, 42, 53]，但是近期通过 Fisher Vector 编码（FV）[54~56]已经成为比较首选的做法，还有一种 FV 的简化版本为局部聚合描述符向量（VLAD）[57]，在 Jain M 等人[58~61]中得到了成功的应用；第二种使用时空词典学习和稀疏编码进行聚合，代表性的工作有 Zhu Y 等人[62~65]；第三种通过时间一致性进行聚合，即通过将时间信息合并到视频描述符的时空信息中，主要的研究集中在 HMM（隐马尔可夫模型）[66]和 CRF（条件随机场）[67]，代表性的工作有 Hongeng S 等人[68~73]。此类方法较多，可以根据问题的实际情况作出选择，不过第二种和第三种方法目前使用的较多，从整体效果来看，第三种方法一般是最好的。

4　深度学习方法

相比于传统方法，深度学习方法不用人工主动去提取特征，保留了视频中更多有价值的信息，从效果来讲，一般优于传统方法。深度学习方法应用在人体行为识别不仅要利用到视频的空间信息，还要用到视频的时间信息，这也是该方法研究的重点。

4.1　时空网络

时空网络重点在于如何提取视中的时间信息，一般使用 CNN 提取空间特征，再利用其他如 LSTM 等方法提取时间信息，时间信息和空间信息使用的类似于电路中的串联架构，这种网络架构在早期的方法中比较流行，效果一般也优于传统的方法，得到了广泛的应用。Li C 等人[74]提出了基于 LSTM 和 CNN 的方法，提取多个人工定义的不同特征，然后分别输入到 3 个 LSTM 网络和 7 个 CNN 网络共，再把这 10 个网络进行融合，并提出了最大融合、平均融合和逐元素相乘融合三种融合方式，最后输出最终的结果，其中逐元素相乘融合表现最好。

Karpathy 等人[75]在卷积网络中提出了晚融合、早融合和慢融合以使网络一次能够连续输入多帧，如图 5 所示，这样能够获取视频中的时间信息，再通过一个 CNN 网络进行处理，类似还有 Chen 等人[61~63]的工作。Donahue J 等人[76]提出了 LRCN，首先通过 CNN 提取空间信息，再经过一个 LSTM 网络提取视频中的时间信息，最后实现分类。Sun 等人[77]也提出了基于 LSTM 的方法来获取视频序列中的时间信息。

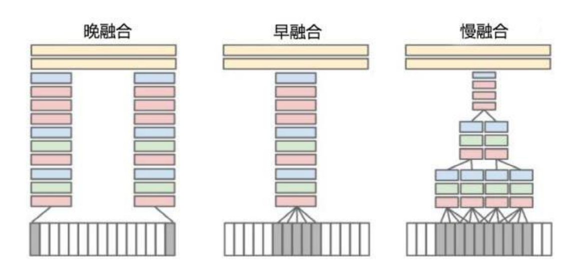


图 5　Karpathy A 等人[75]提出的晚融合、早融合和慢融合策

Ji 等人[78]提出了基于三维 CNN（3D CNN）的方法，3D CNN 是在二维 CNN 基础上加入了时间，因此是三维 CNN，它可以从输入视频中同时学习空间和时间信息，该方法优于二维 CNN 方法。Wang 等人[79]提出了 3D CNN 和 LSTM 相结合的网络，同时对原始视频进行显著性检测，这可以有效降低网络的参数，降低训练的难度，另外对 3D CNN 在 sport-1M 上进行预训练，该方法在 UCF-101 上能够达到 84.0%的识别率。由于 3D CNN 每次只能叠输入固定的帧数，因此不能如 LSTM 一样获取整个视频的时间信息，复杂度也较高，存在一定的局限性，但是效果却比单纯的 CNN 和 LSTM 组合效果要好，说明了在时间维度的卷积能够很好的获取视频中的时间信息，虽然只有固定的帧数。另外，同时使用 3D CNN 和 LSTM 的组合也是一个很好的策略。

4.2　双流网络

双流网络中的时间信息和空间信息采用的方式有点像电路中的并联架构，两个网络开始时互不干涉，各自提取各自的信息，最后采用一定方式进行融合。Simonyan 等人[80]在 2014 年首次提出了创造性的双流网络，如图 6 所示，采用两个相同的 CNN，其中一个网络输入视频帧以获取空间信息，另一个网络输入视频的光流信息以获取时间信息，最后两个网络进行融合，融合方式为平均融合或者使用 SVM 进行融合分类，其中使用 SVM 进行融合分类效果表现最好。与此同时，很多人在双流网络上进行了一系列的改进。Feichtenhofer 等人[81]从融合策略上进行了改进，不像原来双流网络层是最后进行融合，而是从中间就开始进行融合，如图 6 所示，实验表明效果比原有的双流网络好，同时显著减少了参数的数量。Wang 等人[82]使用改进的轨迹（iDT）代替光流提取时间信息，空间网络不变，并将局部 ConvNet 响应汇集在以轨迹为中心的时空管上，其中生成的描述符称为 TDD，最后，使用 Fisher 向量把整个视频中的局部 TDD 聚合成全局超向量，并使用线性 SVM 作为分类器来执

行行为识别。Wang 等人[83]在双流网络的基础上，加入了分段和稀疏化采样的思想，提出了 TSN 网络，如图 7 所示，这样不仅可以减少复杂度，还可以对多个分段进行融合，能够获取更多的上下文信息。时间流网络的输入使用弯曲的光流（warped optical flow fields）来代替原有光流，这样可以消除相机运动带来的影响。另外在训练时加入了交叉形式预训练、正则化、数据增强等技术，使得网络更优。从最终效果分析，时间网络的输入采用光流、轨迹还是光流和轨迹的改进方法，本质上对最终的结果影响不大，决定性的因素还是在网络的结构以及最终的融合方式。

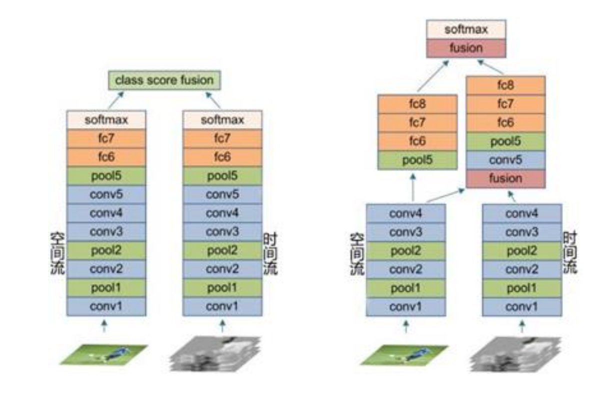


图 6 左图：Simonyan 等人[80]提出的双流网络，采用 RGB 和堆叠的光流帧作为输入；右图：Feichtenhofer 等人[81]的双流融合网络融合策略

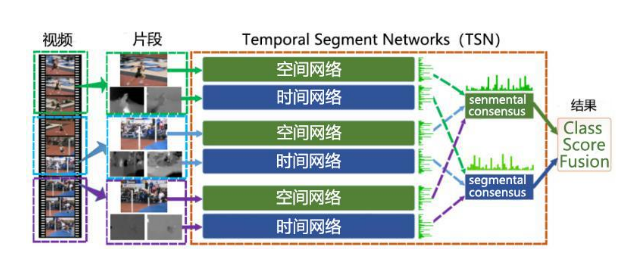


图 7 Wang 等人[82]提出的 TSN，这是划分三段的一个例子

Chen 等人[84]在双流网络的基础上融入半耦合的概念，并应用在极低分辨率的行为识别上。在融合方法上，提出了相加融合、拼接融合和卷积融合，其中卷积融合效果最好。Wang X 等人[85]使用使用了 3D CNN 来代替二维 CNN，为了网络能够支持任意尺寸和和长度的视频输入，在最后一个卷积层不用普通的池化，而是采用 STPP（spatial temporal pyramid pooling），使得输出的特征维度一致。每个网络除了 3D CNN，还引入了 LSTM 或者 CNN-E 来学习时间信息，最后进行融合，其中融合层使用元素最大、元素和、或者级联三种方法，模型如图 8 所示。Gammulle 等人[86]提出了一种双流的 LSTM 网络，开始使用 CNN 提取信息，CNN 是在 ImageNet 预训练的 VGG16，后面再采用 LSTM，并提出了四种融合策略，其中双流的 LSTM 表现最好。Zhao 等人[87]的空间网络采用 3D CNN 网络，时间网络采用了 RNN，RNN 使用了双向的 GRU,输入为人体的骨架序列，在 NTU RGB+D 数据集上取得了比较好的效果。

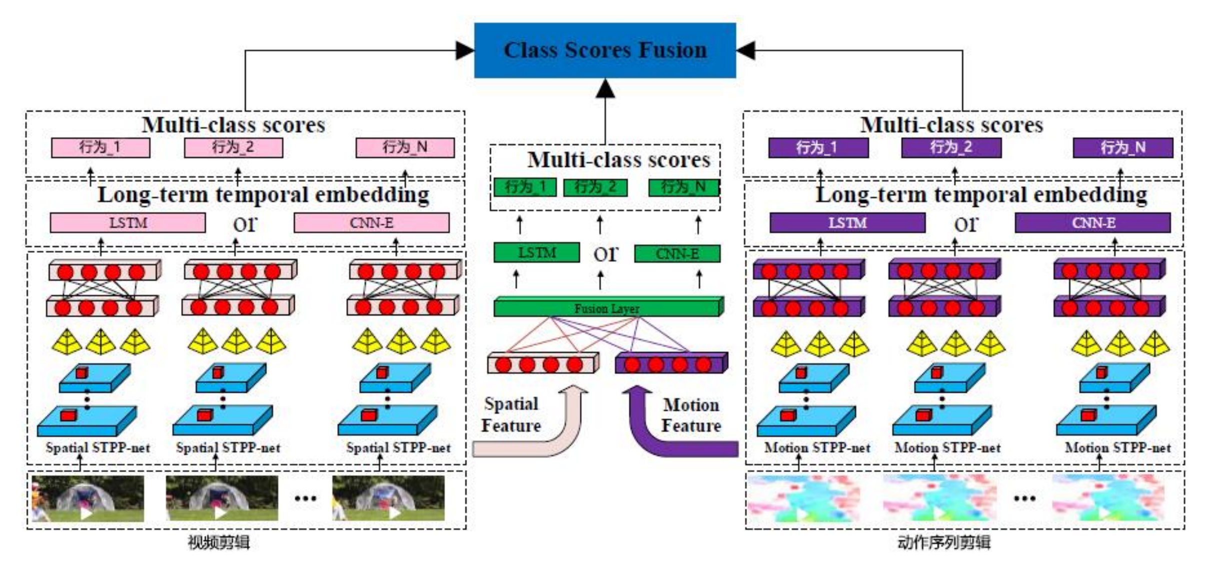


图 8 Wang 等人[85]提出的网络结构,使用了 3D CNN,并使用两种融合策略

总之，双流网络时目前研究最火的框架，不仅仅是由于效果好，还为行为识别的并联架构提供了很好的思路。目前，TSN 网络在 UCF101 上已经达到了最好的结果，可见双流网络的强大，不过同时，深度学习对硬件要求高，极度依赖海量的数据，对实际应用提出了新的挑战。

4.3 其他网络

除了时空网络，还有一些比较优秀的网络架构，特别是一些无监督方法的出现。在视频分析中，注释视频数据的代价高昂，无监督技术显著优于监督的技术。Yan X 等人[88]介绍了 Dynencoder（一种深自动编码器）捕获视频动态，Dynencoder 被证明合成动态纹理是成功的，可以将 Dynencoder 视为表示视频的时空信息的紧凑方式。因此，给定 Dynencoder 的视频的重构误差可以用作分类的均值。Srivastava N 等人[89]提出了 LSTM 自编码器模型（图 9），由编码器 LSTM 和解码器 LSTM 组成，其中编码器 LSTM 接受一个序列作为输入并学习相应的紧凑表示。编码器 LSTM 的状态为序列的紧凑表示，包含序列的外观和动态时间信息，解码器 LSTM 接收学习到的紧凑表示以重建输入序列。

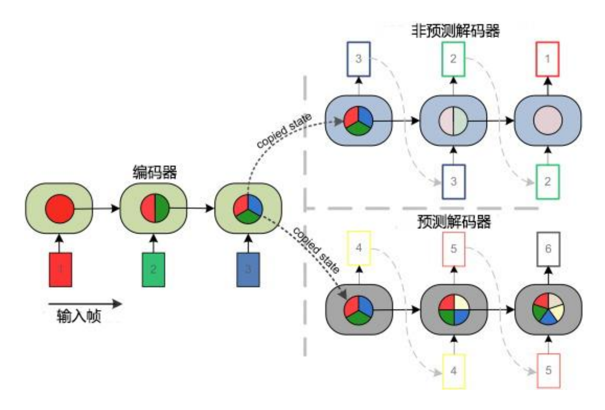


图 9 Srivastava N 等人[89]的 LSTM 自动编码模型，内部的圆圈表示 LSTM 的状态（图中只展示了三帧的情况），非预测解码器尝试以相反的顺序重建原始帧，对预测模型进行预测未来帧 4 的训练; 5 和 6 状态标记上的颜色表示存在来自特定帧的信息

时间相干性是弱监督的一种方法，如果模型分别由有序和无序序列馈送为正样本和负样本，则可以通过深度模型学习时间相关性。这个概念已被 Goroshin R 等人[90]和 Wang X 等人[91]用来从无标签的视频中学习特征。Misra I 等人[92]研究如何使用时间相干性来训练行为识别和姿态估计的深度模型（图 10），连体网络[93~95]用元组进行训练，以确定一个给定的序列是否一致。另一个与时间相关性相关的研究是 Wang 等人[96]的行为识别，分为两个阶段进行，如图 11 所示。对于视频帧集合 $\{x_1,x_2\cdots,x_n\}$，分为先决条件集 X_p（式（1））和效果集 X_e（式（2））分别输入相同的网络，然后通过从 X_p 提取的高级描述符映射到从 X_e 提取的高级描述符所需的转换来标志动作。

$$X_p=\{x_1,x_2\cdots,x_p\} \tag{1}$$

$$X_e=\{x_e,x_{e+1}\cdots,x_n\} \tag{2}$$

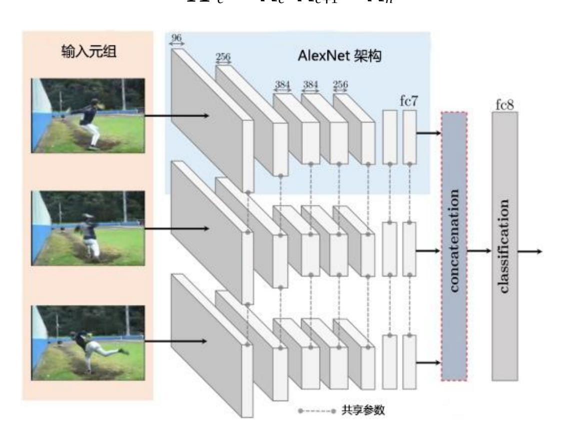


图 10 Misra I 等人[92]使用的连体三重网络

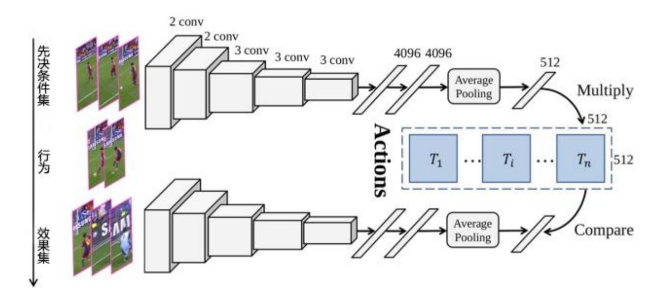


图 11 Wang 等人[96]的平行卷积结构用于提取前置和后置特征

目前，无监督和弱监督的方法也受到广泛的青睐，由于无需人工标签或者少量标签，具有很大的应用价值，是未来的非常有前景的一个研究方向之一。但是由于效果还没有受监督的方法好，因此还有很长的一段路要走。

5 结束语

本文系统地讲解了人体行为识别相关领域的数据集和方法，包含传统的方法和近期比较流行深度学习方法，目前深度学习方法已经成为了主流趋势，并从简单的模型向着复杂模型演化，从最初的监督方法到弱监督方法及以后的无监督方法，都是未来发展的趋势。动作分割的引入，使得连续的行为识别成为可能，但是目前的动作分割算法精度还很低，远远达不到应用的要求，未来还有很长的一段路要走，动作分割和行为识别融合在一起也是未来的发展趋势。

参考文献：

[1] 朱红蕾，朱昶胜，徐志刚. 人体行为识别数据集研究进展 [J/OL]. 自动化学报：1-27. (2018-04-30) . https://doi. org/10. 16383/j. aas. 2018. c170043. (Zhu Honglei, Zhu Changsheng, Xu Zhigang, Research advances on human activity recognition datasets [J/OL]. ACTA AUTOMATICA SINICA: 1-27. (2018-04-30) .)

[2] Schuldt C, Laptev I, Caputo B. Recognizing human actions: a local SVM approach [C]// Proc of the 17th International Conference on Pattern Recognition. 2004: 32-36.

[3] Blank M, Gorelick L, Shechtman E, *et al.* Actions as space-time shapes [C]// Proc of the 10th IEEE International Conference on Computer Vision. 2005: 1395-1402.

[4] Gorelick L, Blank M, Shechtman E, *et al.* Actions as space-time shapes [J]. IEEE transactions on pattern analysis and machine intelligence, 2007, 29 (12): 2247-2253.

[5] Weinland D, Ronfard R, Boyer E. Free viewpoint action recognition using motion history volumes [J]. Computer vision and image understanding, 2006, 104 (2-3): 249-257.

[6] Laptev I, Marszalek M, Schmid C, *et al.* Learning realistic human actions from movies [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. 2008: 1-8.

[7] Marszalek M, Laptev I, Schmid C. Actions in context [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. 2009: 2929-2936.

[8] Rodriguez M D, Ahmed J, Shah M. Action mach a spatio-temporal maximum average correlation height filter for action recognition [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. 2008: 1-8.

[9] Liu Jingen, Luo Jiebo, Shah M. Recognizing realistic actions from videos"in the wild" [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. 2009: 1996-2003.

[10] Reddy K K, Shah M. Recognizing 50 human action categories of Web videos [J]. Machine Vision and Applications, 2013, 24 (5): 971-981.

[11] Soomro K, Zamir A R, Shah M. UCF101: a dataset of 101 human actions classes from videos in the wild [J]. Computer Science, 2012, 12 (1): 1-7

[12] Kuehne H, Jhuang H, Garrote E, *et al.* HMDB: a large video database for human motion recognition [C]// Proc of IEEE International Conference on Computer Vision. 2011: 2556-2563.

[13] Karpathy A, Toderici G, Shetty S, *et al.* Large-scale video classification with convolutional neural networks [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. 2014: 1725-1732.

[14] CMU graphics lab motion capture database [DB/OL]. (2016-09-27) . http://mocap. cs. cmu. edu, .

[15] Li Wanqing, Zhang Zhengyou, Liu Zicheng. Action recognition based on a bag of 3D points [C]// Proc of IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. Washington DC: IEEE Computer Society, 2010: 9-14.

[16] Wang Jiang, Liu Zicheng, Wu Ying, *et al.* Mining actionlet ensemble for action recognition with depth cameras [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. 2012: 1290-1297.

[17] Ellis C, Masood S Z, Tappen M F, *et al.* Exploring the trade-off between accuracy and observational latency in action recognition [J]. International Journal of Computer Vision, 2013, 101 (3): 420-436.

[18] Wang Jiang, Nie Xiaohan, Xia Yin, *et al.* Cross-view action modeling, learning and recognition [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. 2014: 2649-2656.

[19] Chen Chen, Jafari R, Kehtarnavaz N. Utd-mhad: a multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor [C]// Proc of IEEE International Conference on Image Processing. 2015: 168-172.

[20] Shahroudy A, Liu Jun, Ng T T, *et al.* NTU RGB+D: a large scale dataset for 3D human activity analysis [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. 2016. 1010-1019.

[21] Barbič J, Safonova A, Pan J Y, *et al.* Segmenting motion capture data into distinct behaviors [C]// Proc of Graphics Interface. Ontario: Canadian Human-Computer Communications Society, 2004: 185-194.

[22] Zhou Feng, De la Torre F, Hodgins J K. Aligned cluster analysis for temporal segmentation of human motion [C]// Proc of the 8th IEEE International Conference on Automatic Face & Gesture Recognition. 2008: 1-7.

[23] Zhou Feng, De la Torre F, Hodgins J K. Hierarchical aligned cluster analysis for temporal clustering of human motion [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2013, 35 (3): 582-596.

[24] Xia Guiyu, Sun Huaijiang, Feng Lei, *et al.* Human motion segmentation via robust kernel sparse subspace clustering [J]. IEEE Trans on Image Processing, 2018, 27 (1): 135-150.

[25] Lea C, Reiter A, Vidal R, *et al.* Segmental spatiotemporal cnns for fine-grained action segmentation [C]// Proc of European Conference on Computer Vision. Cham: Springer, 2016: 36-52.

[26] Lea C, Flynn M D, Vidal R, *et al.* Temporal convolutional networks for action segmentation and detection [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. Washington DC: IEEE Computer Society, 2017: 156-165.

[27] Devanne M, Berretti S, Pala P, *et al.* Motion segment decomposition of RGB-D sequences for human behavior understanding [J]. Pattern Recognition, 2017, 61 (1): 222-233.

[28] Vögele A, Krüger B, Klein R. Efficient unsupervised temporal segmentation of human motion [C]// Proc of ACM SIGGRAPH//Eurographics Symposium on Computer Animation. Copenhagen: Eurographics Association, 2014: 167-176.

[29] Borzeshi E Z, Concha O P, Da Xu R Y, *et al.* Joint Action Segmentation and Classification by an Extended Hidden Markov Model [J]. IEEE Signal Processing Letters, 2013, 20 (12): 1207-1210.

[30] Liu Shenglan, Feng Lin, Liu Yang, *et al.* Manifold warp segmentation of human action [J]. IEEE Trans on Neural Networks and Learning Systems, 2018, 29 (5): 1414-1426.

[31] Bobick, A. F, Davis, J. W. The recognition of human movement using temporal templates [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2001, 23 (3): 257-267.

[32] Blank M, Gorelick L, Shechtman E, *et al.* Actions as space-time shapes [C]// Proc of the 10th IEEE International Conference on Computer Vision. 2005: 1395-1402.

[33] Weinland D, Ronfard R, Boyer E. Free viewpoint action recognition using motion history volumes [J]. Computer Vision and Image Understanding, 2006, 104 (2): 249-257.

[34] Yilmaz A, Shah M. Actions sketch: a novel action representation [C]// Proc of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005: 984-989.

[35] Dollar P, Rabaud V, Cottrell G, *et al.* Behavior recognition via sparse spatio-temporal features [C]// Proc of Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance. 2006: 65-72.

[36] Matikainen P, Hebert M, Sukthankar R. Trajectons: action recognition through the motion analysis of tracked features [C]// Proc of IEEE International Conference on Computer Vision. 2009: 514-521.

[37] Laptev I. On space-time interest points [J]. International Journal of Computer Vision, 2005, 64 (2-3): 10-123.

[38] Harris C. A combined corner and edge detector [C]// Proc of Alvey Vision Conference. Mancheste: Alvey Vision Club, 1988: 147-151.

[39] Liu Jingen, Luo Jiebo, Shah M. Recognizing realistic actions from videos“in the wild” [C]// Proc of IEEE International Conference on Computer Vision and Pattern Recognition. 2009: 1996-2003.

[40] Klaser A, Marszałek M, Schmid C. A spatio-temporal descriptor based on

3d-gradients [C]// Proc of the 19th British Machine Vision Conference. Leeds: BMVC Press, 2008: 275: 1-10.

[41] Dalal N, Triggs B. Histograms of oriented gradients for human detection [C]// Proc of IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Washington DC: IEEE Computer Society, 2005: 886-893.

[42] Laptev I, Marszalek M, Schmid C, *et al.* Learning realistic human actions from movies [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. 2008: 1-8.

[43] Dalal N, Triggs B, Schmid C. Human detection using oriented histograms of flow and appearance [C]// Proc of the European Conference on Computer Vision. Berlin: Springer, 2006: 428-441.

[44] Ojala T, Pietikäinen M, Mäenpää T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns [J]. IEEE Trans on Pattern Analysis & Machine Intelligence, 2000, 24 (7): 971-987.

[45] Zhao Guoying, Pietikainen M. Dynamic texture recognition using local binary patterns with an application to facial expressions [J]. IEEE Trans on Pattern Analysis & Machine Intelligence, 2007, 29 (6): 915-928.

[46] Kellokumpu V, Zhao Guoying, Pietikäinen M, *et al.* Human activity recognition using a dynamic texture based method [C]// Proc of the British Machine Vision Conference. Leeds: BMVC Press, 2008: 88. 1-88. 10.

[47] Norouznezhad E, Harandi M T, Bigdeli A, *et al.* Directional space-time oriented gradients for 3d visual pattern analysis [C]// Proc of the European Conference on Computer Vision. Berlin: Springer, 2012: 736-749.

[48] Schuldt C, Laptev I, Caputo B. Recognizing human actions: a local SVM approach [C]// Proc of International Conference on Pattern Recognition. 2004: 32-36.

[49] Sanin A, Sanderson C, Harandi M T, *et al.* Spatio-temporal covariance descriptors for action and gesture recognition [C]// Proc of IEEE Workshop on Applications of Computer Vision. 2013: 103-110.

[50] Jiang YuGang, Dai Qi, Xue Xiangyang, *et al.* Trajectory-based modeling of human actions with motion reference points [C]// Proc of the European Conference on Computer Vision. Florence: Springer-Verlag, 2012: 425-438.

[51] Wang Heng, Schmid C. Action recognition with improved trajectories [C]// Proc of IEEE International Conference on Computer Vision. 2014: 3551-3558.

[52] Wang Heng, Ullah M M, Klaser A, *et al.* Evaluation of local spatio-temporal features for action recognition [C]// Proc of the British Machine Vision Conference. London: BMVC Press, 2009: 124. 1-124. 11.

[53] Kovashka A, Grauman K. Learning a hierarchy of discriminative space-time neighborhood features for human action recognition [C]// Proc of Computer Vision and Pattern Recognition. 2010: 2046-2053.

[54] Dan O, Verbeek J, Schmid C. Action and event recognition with fisher vectors on a compact feature set [C]// Proc of IEEE International Conference on Computer Vision. Washington DC: IEEE Computer Society, 2013: 1817-1824.

[55] Peng Xiaojiang, Zou Changqing, Qiao Yu, *et al.* Action recognition with stacked fisher vectors [C]// Proc of the European Conference on Computer Vision. Cham: Springer, 2014: 581-595.

[56] Wang Heng, Kläser A, Schmid C, *et al.* Action recognition by dense trajectories [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. Washington DC: IEEE Computer Society, 2011: 3169-3176.

[57] Arandjelovic R, Zisserman A. All about VLAD [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. 2013: 1578-1585.

[58] Jain M, Jegou H, Bouthemy P. Better exploiting motion for better action recognition [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. 2013: 2555-2562.

[59] Xing Dong, Wang Xianzhong, Lu Hongtao. Action recognition using hybrid feature descriptor and VLAD video encoding [C]// Proc of IEEE Conference on Computer Vision. Singapore: Springer, 2014: 99-112.

[60] Kantorov V, Laptev I. Efficient feature extraction, encoding, and classification for action recognition [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. Washington DC: IEEE Computer Society, 2014: 2593-2600.

[61] Chen Sun, Nevatia R. Large-scale Web video event classification by use of Fisher Vectors [C]// Proc of IEEE Workshop on Applications of Computer Vision. Washington DC: IEEE Computer Society, 2013: 15-22.

[62] Zhu Yan, Zhao Xu, Fu Yuncai, *et al.* Sparse coding on local spatial-temporal volumes for human action recognition [C]// Proc of the Asian Conference on Computer Vision. Queenstown. Springer, 2011: 660-671.

[63] Guha T, Ward R K. Learning sparse representations for human action recognition [J]. IEEE Trans on Pattern Analysis & Machine Intelligence, 2012, 34 (8): 1576.

[64] Somasundaram G, Cherian A, Morellas V, *et al.* Action recognition using global spatio-temporal features derived from sparse representations [J]. Computer Vision & Image Understanding, 2014, 123 (7): 1-13.

[65] Sadanand S, Corso J J. Action bank: a high-level representation of activity in video [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. 2012: 1234-1241.

[66] Rabiner L R. A tutorial on hidden Markov models and selected applications in speech recognition [M]// Readings in Speech Recognition. 1990: 267-296.

[67] Lafferty J D, Mccallum A, Pereira F C N. Conditional random fields: probabilistic models for segmenting and labeling sequence data [C]// Proc of the 8th International Conference on Machine Learning. San Francisco: Morgan Kaufmann, 2001: 282-289.

[68] Hongeng S, Nevatia R. Large-Scale Event detection using semi-hidden Markov models [C]// Proc of IEEE International Conference on Computer Vision. Washington DC: IEEE Computer Society, 2003: 1455.

[69] Koller D, Tang K, Li F F. Learning latent temporal structure for complex event detection [C]// Proc of Computer Vision and Pattern Recognition. 2012: 1250-1257.

[70] Sun C, Nevatia R. ACTIVE: activity concept transitions in video event classification [C]// Proc of IEEE International Conference on Computer Vision. 2013: 913-920.

[71] Quattoni A, Wang S, Morency L P, *et al.* Hidden conditional random fields [J]. IEEE Trans on Pattern Analysis & Machine Intelligence, 2007, 29 (10): 1848-1853.

[72] Wang Yang, Mori G. Hidden part models for human action recognition: probabilistic versus max margin [J]. IEEE Trans on Pattern Analysis & Machine Intelligence, 2011, 33 (7): 1310-23.

[73] Song Yale, Morency L P, Davis R. Action recognition by hierarchical sequence summarization [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. Washington DC: IEEE Computer Society, 2013: 3562-3569.

[74] Li Chuankun, Wang Pichao, Wang Shuang, *et al.* Skeleton-based action recognition using LSTM and CNN [C]// Proc of IEEE International Conference on Multimedia & Expo. 2017: 585-590.

[75] Karpathy A, Toderici G, Shetty S, *et al.* Large-scale video classification with convolutional neural networks [C]// Proc of the IEEE Conference on Computer Vision and Pattern Recognition. 2014: 1725-1732.

[76] Donahue J, Hendricks L A, Guadarrama S, *et al.* Long-term recurrent convolutional networks for visual recognition and description [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. 2015: 677-691.

[77] Sun Lin, Jia Kui, Yeung D Y, *et al.* Human action recognition using factorized spatio-temporal convolutional networks [C]// Proc of IEEE International Conference on Computer Vision. Washington DC: IEEE Computer Society, 2015: 4597-4605.

[78] Ji Shuiwang, Yang Ming, Yu Kai. 3D convolutional neural networks for human action recognition [J]. IEEE Trans on Pattern Analysis & Machine Intelligence, 2013, 35 (1): 221.

[79] Wang Xuanhan, Gao Lianli, Song Jingkuan, *et al.* Beyond frame-level cnn: Saliency-aware 3D CNN with lstm for video action recognition [J]. IEEE Signal Processing Letters, 2017, 24 (4): 510-514.

[80] Simonyan K, Zisserman A. Two-stream convolutional networks for action recognition in videos [C]// Advances in Neural Information Processing Systems. Montreal: NIPS Press, 2014: 568-576.

[81] Feichtenhofer C, Pinz A, Zisserman A. Convolutional Two-Stream Network Fusion for Video Action Recognition [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. 2016: 1933-1941.

[82] Wang Limin, Qiao Yu, Tang Xiaoou. Action recognition with trajectory-pooled deep-convolutional descriptors [C]// Proc of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 4305-4314.

[83] Wang Limin, Xiong Yuanjun, Wang Zhe, *et al.* Temporal segment networks: Towards good practices for deep action recognition [C]// Proc of the European Conference on Computer Vision. Cham: Springer, 2016: 20-36.

[84] Chen Jiawei, Wu J, Konrad J, *et al.* Semi-coupled two-stream fusion convnets for action recognition at extremely low resolutions [C]// Proc of IEEE Winter Conference on Applications of Computer Vision. 2017: 139-147.

[85] Wang Xuanhan, Gao Lianli, Wang Peng, *et al.* Two-stream 3-D convNet fusion for action recognition in videos with arbitrary size and length [J]. IEEE Trans on Multimedia, 2018, 20 (3): 634-644.

[86] Gammulle H, Denman S, Sridharan S, *et al.* Two stream LSTM: a deep fusion framework for human action recognition [C]// Proc of IEEE Winter Conference on Applications of Computer Vision. 2017: 177-186.

[87] Zhao Rui, Ali H, van der Smagt P. Two-stream RNN//CNN for action recognition in 3D videos [J]. Intelligent Robots and Systems, 2017, 10 (1): 4260-4267.

[88] Yan Xing, Chang Hong, Shan Shiguang, *et al.* Modeling video dynamics with deep dynencoder [C]// Proc of the European Conference on Computer Vision. Cham: Springer, 2014: 215-230.

[89] Srivastava N, Mansimov E, Salakhudinov R. Unsupervised learning of video representations using lstms [C]// Proc of International Conference on Machine Learning. 2015: 843-852.

[90] Goroshin R, Bruna J, Tompson J, *et al.* Unsupervised learning of spatiotemporally coherent metrics [C]// Proc of IEEE International Conference on Computer Vision. 2015: 4086-4093.

[91] Wang Xiaolong, Gupta A. Unsupervised learning of visual representations using videos [C]// Proc of IEEE International Conference on Computer Vision. Washington DC: IEEE Computer Society, 2015: 2794-2802.

[92] Misra I, Zitnick C L, Hebert M. Unsupervised learning using sequential verification for action recognition [J]. arXiv: Computer Vision and Pattern Recognition, 2016.

[93] Chopra S, Hadsell R, Lecun Y. Learning a Similarity Metric Discriminatively, with Application to Face Verification [C]// Proc of IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Washington DC: IEEE Computer Society, 2005: 539-546.

[94] Jiwen Lu, Junlin Hu, YapPeng Tan. Nonlinear metric learning for visual tracking [J]. IEEE Trans on Circuits and Systems for Video Technology, 2016, 26 (11): 2056-2068.

[95] Varior R R, Shuai B, Lu J, *et al.* A siamese long short-term memory architecture for human re-identification [C]// Proc of the European Conference on Computer Vision. Cham: Springer, 2016: 135-153.

[96] Wang Xiaolong, Farhadi A, Gupta A. Actions~transformations [C]// Proc of IEEE International Conference on Computer Vision and Pattern Recognition. 2016: 2658-2667.